《心理科学进展》录用版

基于过程数据的问题解决能力测量及数据分析方法*

刘耀辉[1]　徐慧颖[1]　陈琦鹏[1]　詹沛达[1,2]

([1]浙江师范大学教师教育学院心理学系，金华 321004)

([2]浙江省智能教育技术与应用重点实验室, 金华 321004)

摘　要　问题解决能力是指在没有明显解决方法的情况下个体从事认知加工以理解和解决问题情境的能力。对问题解决能力的测量需要借助相对更复杂、更真实、具有可交互性的问题情境来诱导问题解决行为的呈现。使用虚拟测评抓取问题解决的过程数据并分析其中所蕴含的潜在信息是当前心理计量学中测量问题解决能力的新趋势。首先，回顾问题解决能力测量方式的发展：从纸笔测验到虚拟测评。然后，总结对比两类过程数据的分析方法：统计建模法和数据挖掘法。最后，从非认知因素的影响、多模态数据的利用、问题解决能力发展的测量、其他高阶思维能力的测量和问题解决能力概念及结构的界定五个方面展望未来可能的研究方向。

关键词　问题解决能力，过程数据，虚拟测评，计算机化测验，高阶思维能力

1　引言

“在现代社会里，所有生活都是问题解决(In modern societies, all of life is problem solving)”(p.13, OECD, 2014)。Mayer (1990)将问题解决(problem solving)定义为在没有明显解决方法的情况下，将一个给定情境转换为目标情境的认知加工过程。基于此，OECD (2013)将问题解决能力(problem-solving competence)[1]定义为在没有明显解决方法的情况下个体从事认知加工以理解和解决问题情境的能力；同时包括个体参与问题解决的意愿。其中，认知加工可进一步细分为(1)探索和理解(exploring and understanding)、(2)表征和构想(representing and formulating)、(3)计划和执行(planning and executing)和(4)监测和反思

收稿日期：2021-07-08
*国家自然科学基金青年科学基金项目(31900795)资助。
通讯作者：詹沛达, E-mail: pdzhan@gmail.com

[1] 不同研究对“competence”一词的翻译存在差异，其译文包括但不限于“能力”、“素养”和“胜任力”；针对问题解决这一特定主题，作者认为将“competence”译为“能力”更合适也更通俗易懂。但同时也请读者留意其与“ability”和“capacity”等词的差异性。

(monitoring and reflecting)。我国也于 2016 年发布的《中国学生发展核心素养》明确将问题解决作为实践创新的基本点之一，强调学生要"善于发现和提出问题，有解决问题的兴趣和热情；能依据特定情境和具体条件，选择制订合理的解决方案；具有在复杂环境中行动的能力等"。

区别于利用特定领域或问题情境的专业知识或技能的能力，问题解决能力聚焦于能处理真实生活中所遇问题的认知技能，其包括在环境中获取和使用新知识的能力或采用新方式结合个体已有的知识去解决新问题的能力。作为一种不局限于特定问题(任务)情境的一般化能力，问题解决能力所涉及的内容远不止对个体所积累的知识的再现，它还涉及到对认知和实践技能、创造力和其他社会心理资源(比如态度、动机和价值观)的调动(OECD, 2013)。另外，OECD (2013)对问题解决能力的定义强调个体在解决问题时的认知加工过程，并明确指出"学生对评估题目的作答——他们的探索策略，在建模问题时使用的表征，数字和非数字答案，或对问题如何解决的扩展解释——将用于推断他们所采用的认知加工过程"(OECD, 2013, p.122)。

问题解决能力作为一种重要的高阶思维能力[2](Autor & Dorn, 2009)，是个体适应社会与生活的必备特质，也是个体胜任未来工作的核心能力之一。换句话说，具有高水平问题解决能力的人才是促进新时代社会进步的主要动力。然而，对问题解决能力的测量需要依托于真实的、复杂的、具有可交互性的问题情境(任务)，以充分展现问题解决的过程并保证测量的效度；因此，如何实现对个体问题解决能力的客观测量不仅对传统的心理测量方式(例如，采用诸如李克特式题目的纸笔测验)提出了挑战，也对传统的心理测量数据分析方法和理论(例如，经典测量理论(classical test theory, CTT)和题目作答理论(item response theory, IRT))提出了挑战。

面对信息智能时代的全新挑战，提升高阶思维能力、落实核心素养，并建构与之相应的新测评体系显得尤为迫切。近些年，随着心理与教育测量理论与应用研究的发展，尤其是近两年受新冠肺炎(COVID-19)疫情的影响，计算机(网络)化测评形式逐渐成为人们的关注焦点和现实需求。虚拟测评(virtual assessment)是指在计算机化虚拟环境中进行的，可利用虚拟环境特性的测评方式(Agard & von Davier, 2018)，常见的有情景化(scenario-based)、

[2] 高阶思维是指发生在较高层次水平上的认知活动，包括批判性思维、创造性思维、问题解决和决策等，其不仅影响着个体在学业或事业上的表现，也是当代社会发展对人才的基本要求(钟志贤, 2004; Brookhart, 2010; Carroll & Harris, 2020)。

模拟化(simulation-based)和游戏化(game-based)测评。虚拟测评是对传统测评的革新，它更具真实性、情景性和趣味性，能够增加学生的代入感、公平感并缓解测验焦虑，进而促使学生展现出“真实的自己”(Banfield & Wilkerson, 2014; Li et al., 2015)。使用虚拟测评探究学生高阶思维能力或学科核心素养已成为心理与教育测量的新趋势(Liu et al., 2018; Shute & Moore, 2018; 孙鑫等, 2018; 袁建林, 刘红云, 2017)。比如，徐俊怡和李中权(2021)对游戏化测评的概念、范式和实践应用做了详细的阐述；孙鑫等人(2018)和 Shute 和 Rahimi (2020)采用游戏化测评分别测量了学生的推理能力和创造力。除带有实验设计色彩的小规模测评外，诸如国际学生评估项目(Programme for International Student Assessment, PISA)和美国教育进步测评(National Assessment of Educational Progress, NEAP)等大规模测评项目也已经开始使用虚拟测评工具来测量学生的高阶思维能力(OECD, 2016; NCES, 2014)。比如，PISA 2012 和 NEAP 2014 探究了学生的个体问题解决能力；PISA 2015 探究了学生的合作问题解决能力；我国国家基础教育质量监测也于 2020 年开始使用虚拟测评工具测量学生的科学探究能力。

与传统测评方式相比，虚拟测评可基于日志文件(log-file)同时抓取个体作答的结果数据(outcome data)和过程数据(process data)。结果数据是指诸如题目作答精度等传统数据；而过程数据是指带有时间戳(time stamp)的能够反映个体解决问题过程的人机或人人交互数据(Bergner & von Davier, 2018; Hao, Shu, & von Davier, 2015)，包括题目层面过程数据(例如，题目作答时间、题目操作(鼠标点击)次数和答案修改(试错)次数)和相对更为精细的操作层面过程数据(例如，操作历程、操作时间)。分析过程数据有助于了解个体的问题解决过程、探究个体的问题解决策略，对精准诊断学习现状、促进学习发展具有重要作用(Bergner et al., 2018; Jiao et al., 2019; 袁建林, 刘红云, 2020)。对过程数据的分析使得研究重点从探究“结果是什么”转变为探究“结果是如何产生”(Greiff et al., 2015)。与关注结果数据的传统测评相比，额外关注过程数据的虚拟测评对传统的测评数据分析方法提出了挑战。如何合理地分析与利用过程数据，已成为当前心理与教育测量学、教育数据挖掘和学习分析等交叉学科领域的研究新热点与难点。

综上所述，作为一种高阶思维能力，问题解决能力的测量与传统心理特质的测量存在较大差异：前者需要借助相对更复杂、更真实、具有可交互性的问题情境来诱导问题解决

行为(过程)的呈现。换句话说，反映问题解决能力的行为样本相比于反映传统心理特质的更为复杂。这对问题解决能力的测量方式和相应的数据分析方法都带来了挑战。为回答如何客观、准确地测量个体的问题解决能力，以及如何科学、合理地分析虚拟测评中的过程数据这两个问题，如图 1 所示，本文将围绕问题解决能力的测量及数据分析方法这一主题，从(1)问题解决能力测量方式的发展以及(2)过程性数据分析方法两个方面展开阐述，并从非认知因素的影响、多模态数据的利用、问题解决能力的发展和其他高阶思维能力的测量四个方面展望未来可能的研究方向，以期为国内学者更全面地了解问题解决能力的测量及为过程性数据的分析方法提供理论参考。

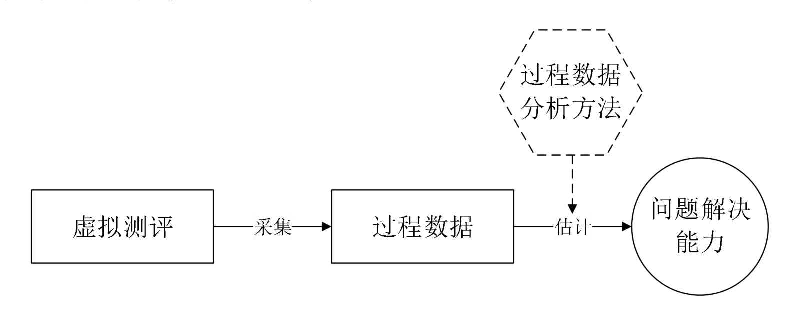


图 1. 基于虚拟测评中过程数据的问题解决能力测量.

2　问题解决能力测量方式的发展

2.1　早期问题解决能力测量方式

由于技术条件的限制，早期问题解决能力测量主要采用传统纸笔测验方式，其特点是基于文字表述给个体营造一定的问题情境，常见于各学科领域内的学业成就测验。Novak (1961)认为问题解决能力的测试应该允许被试在多个选项中选择其认为最正确的答案，同时对于被试的每一步选择，都应当给予反馈。基于此，Novak 将作答环节分成三部分(如图 2)，每一部分提供给被试两个选择，被试的选择范围被箭头所限制，但允许被试返回上一部分选择其他选项。该测验过程相当于被试需要在相互关联的三个部分中分别做出选择，且允许被试在不同的作答阶段反思和修改之前的选择(例题见附录图 A1)。最后的得分由专家依据被试提交的最终答案序列给出(例如, 1→2→2 为满分)。

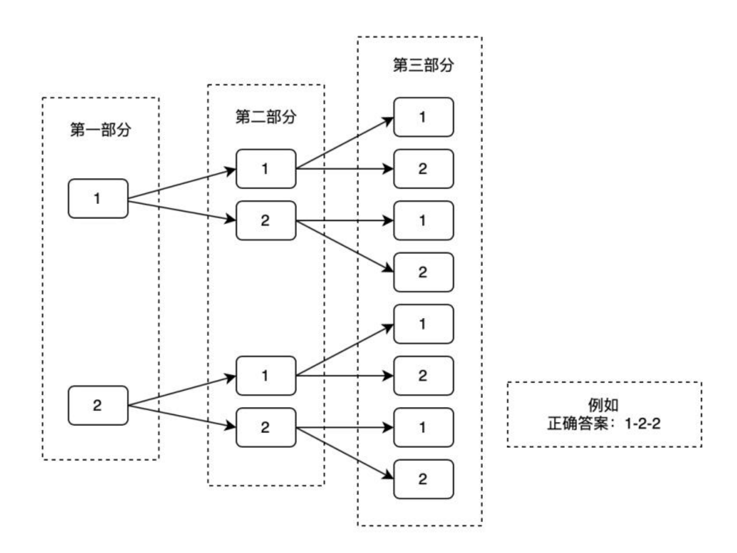


图 2. Novak (1961)提出的分部作答模式

纸笔测验的优点是易于大规模施测、测验工具开发成本较低且对计算机编程技术的依赖程度较低；同时，其缺点也较为明显：一方面是难以记录被试解决问题的详细过程(袁建林, 刘红云, 2020)，另一方面是难以构建真实的、复杂的问题情境。在真实的、复杂的问题情境中往往需要问题解决者与问题情境产生实时交互，这有助问题解决者找出问题产生的原因并做进一步的探索(Greiff et al., 2013)。

另外，值得注意的是，即便是在计算机尚未普及的年代也已出现了小部分虚拟测评。如：模拟经营服装公司的“裁缝店(tailor shop)”和充当消防队长并负责森林消防的“消防队长(fire chief)”系统等(Funke, 1983; Omodei & Wearing, 1995)。对于这些早期的虚拟测评，由于其背后缺乏统一的理论指导框架，导致它们对问题解决能力的测量结果缺乏可比较性(张生等, 2019)。对此，一些心理学家认为在不同领域中有待解决的问题的内容和过程不尽相同，难以提取出有关问题解决能力的全局性理论，应专注于测量不同领域下的问题解决能力(Frensch & Funke, 2002)，如在医疗领域评估被试的病人管理能力和医疗问题解决能力的测评系统(Marshall, 1977; Diserens et al., 1986)。与之不同，另一些持相反观点的心理学家认为通过对问题情境的设置可以构建类似于现实生活中的问题，进而去评估被试的综合问题解决能力。如开发了基于计算机的情景模拟评估系统“洛豪森市(Lohhausen)[3]”，用于分析被试在复杂环境下的高阶思维能力(Doerner, 1980)。

[3] 洛豪森市(Lohhausen)是用计算机模拟现实的一个问题解决评估系统，受试者被要求担任该市“市长”，可以通过调整税率、建立住房等措施来促进城市发展。

21 世纪初，OECD (2003)在前人研究的基础上，勾画了相对全面的问题解决框架(如图 3)。该框架可分为题目设置和问题解决方案生成两部分。在题目设置上，问题情境应贴近个人生活或工作，问题类型需侧重不同的认知过程，同时问题内容也要涉及到不同学科领域的知识。在问题解决方案生成上，注重学生的内在问题解决过程和推理技能。施测形式上，依然采用了传统的纸笔测验形式，用文字和图片来描述问题情境，并基于每段问题表述设置不同类型的问题，如选择题，简答题等。该框架结合现有理论研究，通过对问题类型的设置，加大了对内在认知过程和推理技能的考量。

整体来看，早期问题解决能力的测量主要采用传统纸笔测验。但由于技术条件的限制，纸笔测验中以文字或图片构建的问题情境相对缺乏真实性和情景性，不具备实时交互功能，难以诱发个体真正的问题解决能力。可以说，面对问题解决能力的测量需求，传统纸笔测验方式已心有余而力不足。对问题解决能力等其他高阶思维能力的测量需求促使测量方式的发展，对个体内在认知过程的重视和对现实问题情境模拟的追求也将提高测量的生态效度。这导致研究者对问题解决能力测量新方式的渴望，而计算机(网络)的高速发展为实现对问题解决能力等其他高阶思维能力的测量带来了希望。

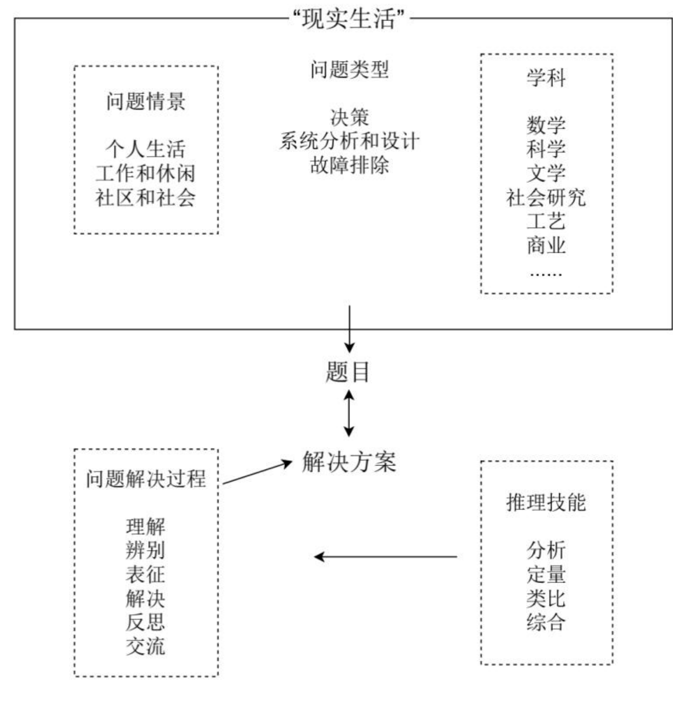


图 3.　OECD (2003)问题解决框架.

2.2　利用虚拟测评测量问题解决能力

以个体为中心的测评应做到在真实情景中对个体的过程性表现进行测评，并给予适当的反馈。比如，Diehl 等人(2005)基于修订的可观察日常生活任务测验(revised observed tasks of daily living)考察老年人的问题解决能力。该测验要求被试在日常生活环境中完成药物使用、电话使用和财务管理等任务，由主试观察并记录下的任务完成情况进行打分。虽然这种基于真实情景的施测方式提高了测验的生态效度，但其施测成本和对主试的高要求阻碍了该测验的大规模的使用。鉴于在真实情景中进行大规模施测缺乏可操作性(例如，成本过高、数据记录不全等)，可在大规模测评中实时并完整抓取个体作答过程数据的虚拟测评逐渐受到人们的关注(Jiao & Lissitz, 2018)。美国心理学会也曾把利用虚拟工具(例如, 游戏)促进学习列入到 2019 年需要关注的 10 个心理学研究趋势之中(Weir, 2018)。

虚拟测评工具的开发是一个相对复杂的过程。相比于传统测评工具(例如, 李克特量表)，虚拟测评工具的开发成本更高、周期更长。因此，程序开发和测验设计等人员在较为统一的工具开发框架下进行及时沟通是必要的。同时，这也有助于保证测量结果之间的可比性。目前，大多数虚拟测评工具是基于证据中心设计(evidence-centered design, ECD; Mislevy et al., 2003)框架开发的(Shute et al., 2017)。该框架认为测量是“基于证据进行推理”的过程，其核心内容是对能力模型、证据模型和任务模型的界定。其中，能力模型界定“测什么”，证据模型界定“怎么测”，任务模型界定“用什么工具测”(如图 4 所示)。另外，还有界定“如何组装测验”的组装模型和“如何呈现任务”的呈现模型，用于测验整体的构建。该框架系统地阐明了复杂测验设计的基本结构、各部分的内涵与功能及相互之间的关系，适用于高阶思维能力或学科核心素养的测评工具开发(袁建林, 刘红云, 2017)。

比如，Zhao 等人(2015)基于 ECD 构建了游戏化测评，用于测量被试的问题解决能力。在能力模型中，从“理解问题给定的条件和约束”、“规划解决方案路径”、“是否有效或高效率地使用工具”和“监测和评估问题解决过程”四个方面去评估被试的问题解决能力。在任务模型中，选用了植物大战僵尸[4]这款游戏作为被试要完成的目标，并设定了相应的任务难度及游戏时长(附录图 A2)。在证据模型的界定中，从可观测的变量中提取了一些行为指标与能力模型建立了联系(附录图 A3)，并用贝叶斯网去搭建各变量之间的数学关系。该游

[4] 植物大战僵尸是一款策略塔防类游戏，玩家需要收集阳光，安置不同的植物，使用其功能以阻挡僵尸的入侵。

戏测评结果与 MircoDYN[5]测试结果相关显著($r = 0.48, p < 0.01$)，基于聚合效度，表明了该游戏化虚拟测评的有效性。

此外，如上文所述，目前诸如 PISA 和 NEAP 等大规模测评项目也已经开始使用虚拟测评工具来测量学生的问题解决能力，比如，PISA 2012 和 NEAP 2014 探究了学生的个体问题解决能力，PISA 2015 探究了学生的合作问题解决能力。以 PISA 2012 的一道题为例(如附录图 A4 所示)，题目呈现了一个 MP3 播放器，学生需通过点击播放器的按钮来了解其工作原理。在此基础上，学生需回答题目对应的 4 个问题。每个问题则侧重考察学生问题解决中不同的认知过程，例如，第一问主要考察学生对题目的探索和理解、第二问主要考察学生问题解决中的计划和执行能力等。该测验通过向学生呈现生活中可能遇到的问题来实现对其问题解决能力的评估，测评结果由系统判定和专家评分两部分组成。同时，大规模的国际化虚拟测评也为各国、各地区之间在人才培养方面提供了参考借鉴的机会。

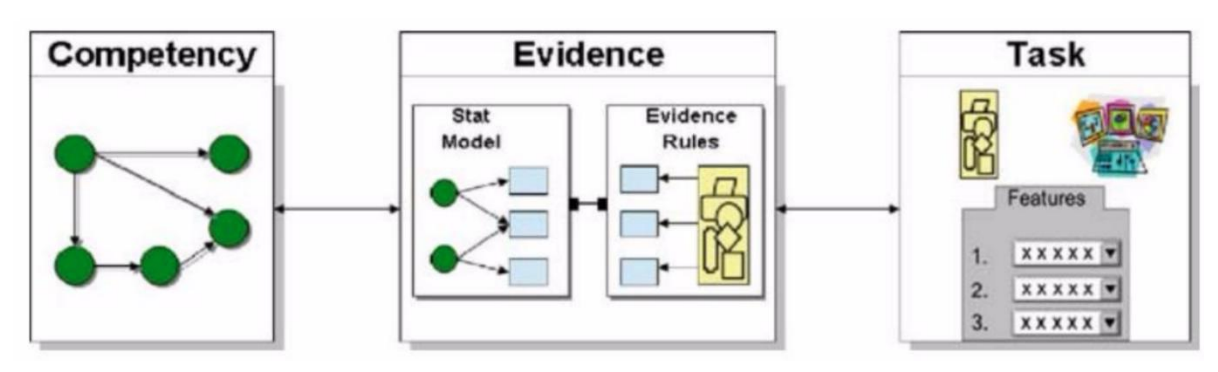


图 4. 证据中心设计框架中的能力模型、证据模型和任务模型 (Zhao et al., 2015).

3　过程性数据分析方法

鉴于虚拟测评的新颖性及过程数据的复杂性，国内外关于过程数据的分析方法的研究均处于起步阶段。经过梳理，大体可将现有的过程数据分析方法分为两类：数据挖掘法(data mining)和统计建模法(statistical modeling)。其中，前者属于探索性研究方法或归纳法，是基于数据驱动的自下而上的研究方法，强调从已有数据入手，对数据进行描述、分析、总结和归纳理论，遵循着“发现的逻辑”；而后者属于验证性方法或演绎法，是基于理论驱动

[5] MircoDYN 是一个基于计算机交互式的动态问题解决评估系统，该系统将多个任务嵌入线性结构方程框架用来评估被试的动态问题解决能力。详细内容可见 Greiff et al. (2012)。

的自上而下的研究方法，强调从理论出发，生成假设，再用数据检验，接受或者拒绝假设，遵循着"证明的逻辑"。如图 5 所示，两种方法的使用形成了一个循环的研究过程(Johnson & Christensen, 2014)，推动着科学研究的发展。

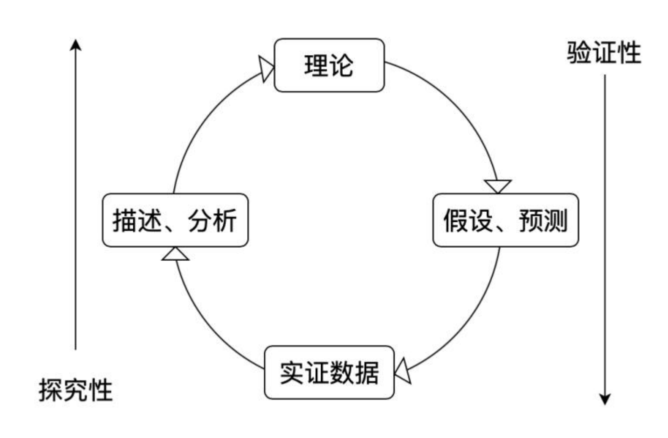


图 5. 循环研究过程 (Johnson & Christensen, 2014)

3.1　数据挖掘法

基于虚拟测评产生的过程数据，因其数据结构的不规则性和来源的复杂性，大幅度增加了分析难度。过程数据背后蕴藏着与问题解决有关的认知加工过程，需要采用特定的方法来挖掘和分析。数据挖掘是指从大量数据中通过算法来揭示有意义的新的关系、趋势和模式的过程(王光宏, 蒋平, 2004)，是"从数据中发现知识规律(knowledge discovery in databases)"(Fayyad et al., 1996)的过程。通过数据挖掘来探索过程数据所蕴含的潜在信息是教育数据挖掘领域的主要研究逻辑，目前主要涉及有监督学习(supervised learning)算法和无监督学习(unsupervised learning)算法这两类。

有监督学习算法是通过已有的训练样本(即已知数据及其对应的类别)来训练分类器(classifier)，再利用训练后的最优分类器将新的研究样本映射到相应的类别中，其中类别总数是已知且固定的。目前，使用有监督学习算法分析过程数据的研究还较少，而决策树(decision tree)是被使用相对较多的一类算法，主要包括分类和回归树(CART; DiCerbo & Kidwai, 2013)及随机森林(random forest; Hao et al., 2016; 孙鑫等, 2018)这两种方法。比如，为了探索可以有效预测被试反映的特征，Han 等人(2019)从被试的行为序列、有策略意义的行为指标和时间纬度三个方面初步提取了 77 个特征变量，通过随机森林和递归特征消

除法筛选出 13 个最有高预测表现的特征变量。例如，是否应用“一次只改变一个变量(vary one thing at a time)”策略和应用该策略的次数这两个特征变量都能有效预测被试是否有较大概率成功解决问题。

无监督学习算法是在事先没有任何训练样本的情况下，直接根据研究样本之间的相似性对样本进行分类，并试图使类内差距最小化且类间差距最大化，其中类别总数是未知且不固定的。目前，在对个体问题解决策略进行分类时，主要采用聚类分析(cluster analysis; Bergner et al., 2014)和自组织映射神经网络(SOM; Soller & Stevens, 2007)。鉴于不同的无监督学习算法可能会给出有差异的分类结果，有研究建议同时使用多种无监督学习算法，比如 Fossey(2017)对比了三种无监督的学习算法，包括 *k*-means、SOM 和使用链接的鲁棒聚类算法(ROCK)；Qiao 和 Jiao(2018)针对同一批数据同时研究了四种有监督学习算法(CART、随机森林、梯度提升决策树和支持向量机)和两种无监督学习算法(*k*-means 和 SOM)的表现。

该方法的优势在于基于训练好的分类器或者不同的聚类规则便可快速实现对个体问题解决能力的分类，同时应用数据挖掘算法也能快速帮助研究者从高维复杂的数据中挖掘出有统计意义的信息，但该方法在心理学领域中的应用和推广还需要理论基础的支撑。一方面就数据挖掘算法而言，其任务是从数据中挖掘隐藏在数据中的模式，刻画当前数据特征或构建高预测率模型(王光宏等, 2004)。但其构建的模型或得出的结论有时并不能给我们带来任何启示，甚至是无用的。对大数据的处理，要注重对其背后含义的理解(吴忭等, 2019)。另一方面，就数据挖掘在心理学中的应用而言，心理学领域注重数据结果的可解释性或其折射出的基于个体或群体的心理过程和潜在特质等。过程数据的价值在于其背后对应的心理过程，单纯的数据驱动无法在跨任务的操作中提取或者构造出可反映个体自身潜在水平的变量(He et al., 2021)，很难得出有效可靠的结论，往往需要专家的进一步介入。比如，在特定情景中，需要专家界定出最优行为序列、判断异常行为或解读数据分析结果(Hao et al.,2015; He et al.,2021; He & von Davier, 2016; Qiao & Jiao, 2018)。另外，对于特定任务下结论的有效性也应持有怀疑的态度。比如，Qiao 和 Jiao (2018)的研究发现所有方法均表现出满意的分类一致性，但在此研究中并未发现时间信息作为分类依据的重要性，这与其他已有研究观点不同(Chen, 2020; Molenaar et al., 2016; Ulitzsch et al., 2021)。再有，在数据的

预处理方面，如数据的筛选、排序、编码等，处理方法也常常因数据类型、分析目的和选用算法的不同而不同；对缺失数据、极端值和重复行为序列的处理也且尚无内在统一标准。

3.2　统计建模法

统计建模法主要是指利用人工建模的思路来分析数据的方法。在统计建模中，一般基于理论假设构建函数模型，同时假设观测变量是由该模型所表达的概率法则随机生成的(洪永淼, 汪寿阳, 2021)。 通过统计模型来解释过程数据所蕴含的潜在信息是心理计量学的主要研究逻辑(Bergner & von Davier, 2018)。符合心理计量学的基本假设：个体的内隐(潜在)特质决定其外显行为。目前，针对记录下的过程数据和结果数据，统计建模法主要包括心理计量联合建模(psychometric joint modeling)、隐马尔可夫建模(hidden Markov modeling)和多水平建模(multilevel modeling)等。

心理计量联合建模是目前最常见的题目层面过程数据分析方法。该方法的逻辑是基于IRT视角下的联合-层级建模框架(joint-hierarchical modeling framework; van der Linden, 2007)，建构针对不同数据源(例如，题目作答结果和题目作答时间)的心理计量模型，然后使用多元正态分布描述多种潜在特质之间的关系。目前，该方法分析的过程数据主要是题目作答时间。基于此，研究者们提出了一系列的联合模型用于探究个体潜在能力、加工速度及两者之间的关系(Fox & Marianti, 2016; Man et al., 2019; Molenaar et al., 2018; Zhan & He, 2021; 詹沛达, 2019)。此外，为满足当前实践对诊断性测评的需求，Zhan 等人 (2018)从认知诊断视角对联合建模框架进行拓广，所提出的联合认知诊断建模框架允许研究者使用不同的高阶认知诊断模型 (e.g., de la Torre & Douglas, 2004)和作答时间模型(van der Linden, 2006)分别作为作答结果和作答时间的测量模型，进而可以同时探究个体的一般学习能力、属性、加工速度及它们之间的关系。

隐马尔可夫建模假设个体的解题历程符合马尔可夫过程并受个体潜在能力的影响，侧重对问题解决过程的建模。其中，个体的解题历程包括外显的操作步骤和内在认知状态的变化(如，问题表征、策略使用)；马尔可夫过程是研究离散事件动态系统状态空间的一种方法，是指在一个随机过程中事物的未来状态仅依赖于当前状态而与过去状态无关。Baker等人(2011)在其研究中验证了马尔可夫过程作为认知模型的可行性，且马尔科夫过程已被广泛应用于过程数据的建模中(Shu et al., 2017)。Molenaar 等人 (2016)把隐马尔可夫模型引

入到联合建模框架中，把个体按特定顺序的作答视为马尔可夫过程，通过分析个体在不同题目上作答时间的变化探究他们个体内(within-subject)加工速度的变化情况。鉴于作答时间可以在一定程度上反映个体对知识的精熟程度，Wang 等人(2018)在认知诊断视角下提出了高阶隐马尔可夫模型，通过分析个体在纵向测验上作答时间的变化测量他们的学习进步情况。实际上，上述两个研究所分析的仍是题目作答时间。与之不同，Shu 等人(2017)针对个体的问题解决过程(操作历程)提出了马尔可夫 IRT 模型，认为个体的当前操作与其上一步操作和其潜在能力有关。该模型把所有可能的相邻操作行为视为操作层面“题目”，进而根据个体在“题目”上的“作答”(例如，是否呈现该操作)去估计其潜在能力。该模型巧妙地将个体的问题解决过程转换为操作层面观察分数，实现了在单题内估计个体潜在能力，为后续研究提供了借鉴和参考。

在传统心理统计中，多水平建模(multilevel modeling)常用于分析因分层抽样导致含有嵌套关系的数据[6]。通过多水平建模可将个体水平上个体数据之间的变异分解为班级、学校或地区等不同水平上的变异，有助剥离出造成个体之间差异的真实原因(刘红云, 骆方, 2008)。Liu 等人(2018)将该逻辑迁移至过程数据分析中，假设由人工赋分得到的操作层面分数嵌套于个体个体水平，并基于该逻辑提出了适用于分析操作历程数据的多水平混合 IRT 模型。该研究与 Shu 等人(2017)类似的是需要先对个体的问题解决历程进行人工赋分；所不同的是该研究把所有可操作项(例如，可选路线)视为操作层面“题目”，把个体的特定操作行为视为操作层面“人”，然后根据“人”在“题目”上的“作答”去估计其潜在能力。鉴于该模型同时包含了 IRT 模型、潜在类别模型和多水平模型的特点，它可在单题内估计个体的问题解决能力并判断其所采用的问题解决策略。

除此之外，近些年也有研究尝试利用题目扩张技术(即将一道虚拟测评题目中正确解答所需的操作流程拆解为多个子流程(或步骤)，并将这些子流程视为相互条件独立的虚假题目(pseudo item)；然后根据个体在解决问题过程中是否呈现出这些子流程，对其进行赋分)，直接使用传统的心理计量模型对过程数据进行分析(Zhan & Qiao, 2020)。这种做法虽然增加了数据预处理的难度，但大幅度降低了数据分析的难度，为分析过程数据提供了新思路。

3.3　两种方法的对比

[6] 通常，多水平数据的分布在个体之间不具备独立性，存在地理距离内、某行政区域内或者特定空间范围内的聚集性(clustering)或相似性。

近些年，在智能时代背景下，研究者们愈发倾向于在技术增强环境(technology-enhanced environment)中探索心理与教育测量的新范式。虚拟测评和数据挖掘技术因其“智能”属性更容易引起研究者和实践者的关注。比如，利用游戏化测评来测量个体的高级认知技能，并采用数据挖掘技术分析数据以实现对个体分类(Qiao & Jiao, 2018)。实际上，数据挖掘技术与潜变量建模在底层逻辑上存在差异：后者主要关注的是隐藏在外显行为数据背后的潜在变量，即假设潜在变量决定外显行为，并通过潜变量模型实现对两者的联接；而前者仅关注外显行为数据的分析，通过计算数据之间的相似性或距离对数据进行分类或聚类。对数据挖掘技术而言，因为不存在理论假设的因果关系，所以我们难以利用其结果来反推导致该结果的原因。因此，数据挖掘技术的结果可解释性通常低于潜变量模型的，而结果的可解释性恰恰是心理与教育测量的重点。

整体而言，采用统计建模法分析过程数据的主要优势是结果的易解释性且符合心理与教育研究的一般过程(如图 6 所示)；其局限性是需要针对不同类型的过程数据分别建模，这也导致目前针对不同类型过程数据的建模逻辑尚未统一。而数据挖掘法的主要优势是可以同时考虑多种过程数据，其局限性是结果的可解释性较差，即无法直接报告个体的具体不足，仍需采用专家判断法做推断。然而，在心理与教育测量中，尤其是在诊断性测量中，结果的易解释性显得尤为重要。另外，现有的数据挖掘方法主要是基于观察变量进行分类，而非基于个体的潜在特质(例如, 认知过程或知识技能)进行分类，在数据源和数据量有限的情况下两种分类结果并不完全等同(Liu & Cheng, 2018)。反观，基于潜在特质进行分类，明确指出个体在特定的认知过程或知识技能上的不足，有助于教师或干预者有针对性地制定补救教学或干预方案。

实际上，数据挖掘法和统计建模法各具优势，在心理与教育测量中，它们适用于解决不同的问题。前者更适用于在具有多变量且不满足特定概率密度函数的复杂数据情境下挖掘隐藏的规律，并依据这些规律对个体进行分类，但同时又不需要解释分类的具体原因的场景。比如，在自适应学习系统中根据学生的学习时长、练习结果、内容偏好等多变量的数据进行分类，进而推荐适合的学习内容，或依据特定评分(级)规则对文字内容(例如，作文)进行自动评分(级)。由于数据挖掘法解决的是分类问题，所以采用该方法的研究常以分类结果来报告个体问题解决能力之间的差异(如，“正确组”、“冗余行为组”、“离群组”等(Qiao

& Jiao, 2018))。相比之下，后者更适用于在满足特定概率密度函数的数据情境下，基于概率密度函数构建可联接外显行为与潜在特质的统计模型，并依据这些统计模型实现对个体潜在特质水平或类别的估计。比如，针对题目作答精度数据，基于 Logistic 函数构建的 IRT 模型，并依据 IRT 模型实现对潜在能力水平的估计；或针对题目作答时间，基于对数正态分布函数构建题目作答时间模型，并依据题目作答时间模型实现对潜在加工速度水平的估计。由于统计建模法以被试参数的形式来反映个体的问题解决能力，所以采用该方法的研究对问题解决能力的报告形式是由被试参数的类型决定的。比如，Shu 等人(2017)用连续潜变量表示个体的问题解决能力，Zhan 和 Qiao (2020)用连续变量表示个体的一般问题解决能力并用类别变量表示个体的问题解决策略。

以基于特定问题拟将个体的问题解决能力分为“高”、“中”和“低”三个类别为例。若采用数据挖掘法，比如有监督学习算法，就需要先采用专家判断法对已知的典型行为数据打标签(如，包含哪些行为表现的数据可以被标记为“高”)，然后将训练数据和对应标签放入分类器进行训练，再用训练好的分类器去分析个体解决该问题时的行为数据，进而实现对个体问题解决能力的分类；而若采用统计建模法，就需要先对观测到的行为数据进行描述性统计，判断其分布形态是否符合某种概率密度函数，然后基于该概率密度函数构建同时包含反映问题解决能力的被试参数和题目参数的统计模型(其中被试参数应为类别变量)，再用所构建的模型去分析个体解决该问题时的行为数据，进而实现对个体问题解决能力的参数估计。

目前，虚拟测评中过程数据的主要作用还是为测量个体的问题解决能力提供信息，仍遵循不可观测的问题解决能力决定可观测的过程数据这一基本假设。鉴于统计建模法可以基于模型预先构建导致外显行为的(理论)原因，更适用于以结果解释为目的应用情境，所以针对问题解决能力测量这一议题，统计建模法仍将发挥主要作用。波普尔指出“不是经验的重复产生心理的信念，而是心理的信念产生经验的重复”(成素梅, 荣小雪, 2003, p. 15)，虽然从已有经验、观测数据中可以归纳出一些有用的结论和概括，但其也仅是提供了一些可能的说法。科学发展的逻辑还须是从理论假设出发，用数据验证理论或者推翻理论，即遵循着“假设检验”的过程和“可证伪原则[7]”。

[7] 可证伪原则是由波普尔提出，其认为科学的理论应具有可证伪性。一个理论的可证伪性就是指该理论推导出的结论在逻辑上或在原则上有可能与一个或一组观察陈述发生抵触。

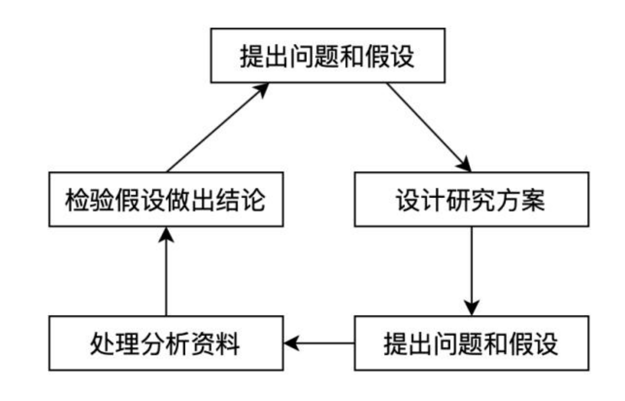


图 6. 心理与教育研究的一般过程

最后，值得注意的是，尽管我们强调基于过程数据的问题解决能力测量及数据分析方法，但国内外相关研究似乎并没有否定根据传统结果数据推断个体问题解决能力的方式，只不过利用过程数据可以更好地呈现出被试解决问题的过程，有助于了解个体呈现不同作答结果的历程，有助于更准确地推断个体的问题解决能力。比如，一气呵成地解决特定问题和经过反复退改地解决同一问题的两位学生，尽管他们的结果是一样的，但他们问题解决能力很可能是不一样的(即似乎前者更高)；而“一气呵成”和“反复退改”正是过程数据所呈现的，单凭结果数据无法区分两位学生的问题解决能力。实际上，无论是统计建模法还是数据挖掘法，都可以联合分析或同时利用结果数据和过程数据。比如，在统计建模法中，心理计量联合建模可以联合分析题目作答精度这一结果数据和题目作答时间这一过程数据；在数据挖掘法中，基于有监督学习算法，可以利用结果数据作为典型行为数据的标签(如，满分作答结果包含哪些必要的行为过程？相比于满分作答结果，得到部分作答结果又缺少了哪几个关键行为过程？)去训练分类器；而基于无监督学习算法，可以利用结果数据来检验分类的有效性(如，探索出的类别是否较好地分离出某个测验中的完成者和未完成者)。

4　讨论与展望

本文对问题解决能力测量方式的发展以及虚拟测评中过程数据的分析方法进行了梳理。测评方式的转变和过程数据的分析对问题解决能力的研究有重要意义，一方面为研究问题解决过程及其影响因素提供了技术的支持，另一方面也有助于实现应用过程数据对其

他高阶思维的测量。但目前在概念界定、数据采集和分析上仍有较大的发展空间，未来研究者可以从非认知因素带来的影响、多模态数据的利用、问题解决能力的发展、其他高阶思维能力的测量和问题解决能力概念及结构的界定五个角度入手，进一步丰富有关问题解决能力的测评研究。

4.1 非认知因素对问题解决能力的影响

李一茗和黎坚(2020)探讨了复杂情境中的问题解决能力的影响因素，认为问题解决能力不仅涉及到推理能力、工作记忆容量和加工速度等认知加工能力，还受到个体的元认知调节、知识背景、动机和情绪等非认知因素的影响。而现有的过程数据分析方法仍主要停留在对特定认知加工过程的建模与分析上。将问题解决能力视为一个笼统的单维潜在特质或仅关注对特定认知加工过程的测量，在测量中忽略了其他非认知因素对问题解决能力的影响。像态度、情感、信念和一些能反应人格特质的非认知因素，被称为非认知能力(祖霁云, Kyllonen, 2019; 徐俊怡, 李中权, 2021)。非认知能力不仅影响着问题解决的过程，也对个体学业和劳动力市场表现有着重要作用(何珺子, 王小军, 2017)。如何扩展现有数据分析方法，纳入对非认知能力的测量与分析，是全面了解个体，提高个体问题解决能力的有效途径。

4.2 利用多模态数据测量问题解决能力

当前对过程数据的挖掘和分析主要集中在题目作答精度、作答时间和行为序列上。这些数据还不足以全面反映个体问题解决中的认知及非认知过程。实际上，虚拟测评等其他计算机化测验的自动化特性使其能够在个体解决问题过程中实时记录不限于过程数据的多种类型数据(即多模态数据)。比如，除题目作答精度和题目作答时间外，通过嵌入式传感器(例如，眼动仪)还可以同步记录诸如眼动和神经活动等生物计量数据(biometric data)。Jeon 等人(2021)通过分析题目作答精度和大脑激活数据，测量了个体的潜在能力和大脑激活水平；Man 和 Harring (2020)通过分析题目作答精度、作答时间和眼动数据，测量了个体的潜在能力、潜在加工速度和潜在专注力水平；Bezirhan 等人(2021)融合分析了题目作答精度、作答时间和重访题目次数(revisit counts, 考生在首次答完某题后再次访问该题的次数)，测量了个体的潜在能力、潜在加工速度和重访题目倾向。另外，Zoanetti (2010)的研究中不仅记录了被试解决问题时的操作信息，同时也搜集了被试的口头表述信息(例如，

被试在某一时刻说："我不明白")和外在行为数据(例如，皱眉、叹气)，进而区分了相似过程数据下的不同认知过程。比如，当不同被试在问题表征阶段花费相似时间时，可结合口头表述信息去判断他们是在理解题目还是在构思解决方案。

在虚拟测评中，由于多模态数据的采集几乎是同时进行的，且它们提供的是有关被试在问题解决时的平行信息，因此，也有研究者将它们称为平行数据(parallel data; Jeon et al., 2021)，比如，被试正确作答某题目耗时 30 秒并投入 20 个视觉注视点。对多模态数据的融合分析，为从更全面的视角理解个体的问题解决能力提供了可能性。未来，随着传感器的可便携性增加及成本下降，多模态数据的采集与分析势必会常态化，非常值得心理与教育测量领域研究者的关注。

4.3　对问题解决能力发展的测量

测量和促进个体能力的发展是心理学与教育学中需要解决的重要问题(Zhan & He, 2021)，对问题解决能力发展变化的研究事关教学设计和教学策略的制定与实施。然而，当前对问题解决能力的测量主要依赖于对横断过程数据的分析，较少依赖于对纵向过程数据的分析。一方面是因为对横断过程数据分析尚未有较为统一的认识和分析范式，探讨可分析纵向过程数据的方法可能还为时尚早；另一方面是因为纵向虚拟测评工具的开发难度更高。

目前，已有一些研究尝试使用虚拟测评中的一些外显指标来评估个体问题解决能力的发展。比如，张博等人(2014)基于游戏化测评使用由成功完成推箱子题目的数量来表示的认知能力、由每题计划时间与作答总时间的比值来表示的元认知能力和由每题所用总步数来表示的认知效率三个指标对比研究了普通儿童和超常儿童的问题解决能力的发展。研究结果表明，11-14 岁之间，超常儿童问题解决能力的发展遵循着高起点，先快后慢的规律；普通儿童则起点较低，发展先慢后快。随着年龄的增长，二者之间差异逐渐缩小。同时，该发展模式也体现在两组儿童的认知能力和元认知能力两个维度上，但在认知效率上，二者之间的差异并没有随施测时间不同而发生显著变化。但值得注意的是这类研究并没有直接对问题解决能力进行估计，进而实现对不同时间点上估计值的发展的测量；因此，对问题解决能力发展的测量仍值得后续研究的关注。

4.4　其他高阶思维能力的测量

如上文所述，除问题解决能力外，高阶思维能力还包括批判性思维能力和创造性思维能力等，其不仅影响着个体在学业或事业上的表现，也是当代社会发展对人才的基本要求(钟志贤, 2004；Brookhart, 2010; Carroll & Harris, 2020)。除问题解决能力外，目前已有很多研究尝试使用虚拟测评去测量诸如创造力、领导力等其他高阶思维能力(Shute & Rahimi, 2020; Stanek & Sabat, 2019)。另外，2022 年 OECD 也计划采用情景化测评方式来测量个体的创造力(OECD, 2019)。未来，随着测量方式及数据分析技术的不断发展，充分利用计算机(网络)技术，尤其是人工智能，并结合便携式和低成本的心理学实验仪器，我们有理由相信可以在大规模测验中实现对高阶思维能力的测量。

4.5　问题解决能力概念及结构的界定

当前国内外对问题解决能力的主要研究基本都是围绕 OECD (2013)对问题解决能力的定义实施的。首先，该定义并没有局限于特定的任务情境；因此，该定义所述的问题解决能力是一种一般化能力或特质。其次，该定义将其所强调的认识加工又进一步细分为(1)探索和理解、(2)表征和构想、(3)计划和执行和(4)监测和反思；同时，值得注意的是，除认知加工外，该定义中还特别强调了个体参与问题解决的意愿。因此，该定义所述的问题解决能力至少具有多维结构，而至于是否满足高阶结构，可能需要后续研究做实证验证或理论阐述。另外，该定义所述的是个体问题解决能力，目前已有研究开始探讨协作问题解决(collaborative problem solving) (如, Unal & Cakir, 2021)；而协作问题解决能力与个体问题解决能力的概念及结构有何区别仍值得后续研究做进一步探讨。最后，OECD (2013)对问题解决能力的定义是否具有跨时代稳健性(即该定义是否会随时代的发展产生变化)也值得后续研究者们的关注。

参考文献

成素梅, 荣小雪. (2003). 波普尔的证伪方法与非充分决定性论题. *自然辩证法研究, 19*(01),15–19+29.

何珺子, 王小军. (2017). 认知能力和非认知能力的教育回报率——基于国际成人能力测评项目的实证研究. *经济与管理研究, 38*(05),66–74.

洪永淼, 汪寿阳. (2021). 大数据、机器学习与统计学:挑战与机遇. *计量经济学报, 1*(01),17–35.

李一茗, 黎坚. (2020). 复杂问题解决能力的概念、影响因素及培养策略. *北京师范大学学报(社会科学版)*, (05),36–48.

刘红云，骆方.(2008).多水平项目反应理论模型在测验发展中的应用. *心理学报*(01), 92-100.

孙鑫, 黎坚, 符植煜. (2018). 利用游戏 log-file 预测学生推理能力和数学成绩——机器学习的应用. *心理学报, 50*(7), 761–770.

王光宏, 蒋平. (2004). 数据挖掘综述. *同济大学学报(自然科学版), 32*(02),246–252.

吴忭, 胡艺龄, 赵玥颖. (2019). 如何使用数据:回归基于理解的深度学习和测评——访国际知名学习科学专家戴维·谢弗. *开放教育研究, 25*(01),4–12.

徐俊怡, 李中权. (2021). 基于游戏的心理测评. *心理科学进展, 29*(03),394–403.

袁建林, 刘红云. (2017). 核心素养测量: 理论依据与实践指向. *教育研究, 38*(07), 21–36.

袁建林, 刘红云. (2020). 过程性测量：教育测量的新范式. *中国考试, 12*,1–9.

詹沛达. (2019). 计算机化多维测验中作答时间和作答精度数据的联合分析. *心理科学, 42*(1), 170–178

张博, 黎坚, 徐楚, 李一茗. (2014). 11～14 岁超常儿童与普通儿童问题解决能力的发展比较. *心理学报, 46*(12),1823–1834.

张生, 任岩, 骆方. (2019). 学生高阶思维能力的评价:复杂问题解决的测量述评. *中国特殊教育, 10*, 90–96.

钟志贤. (2004). 促进学习者高阶思维发展的教学设计假设. *电化教育研究*, (12),21–28.

祖霁云, Patrick Kyllonen. (2019). 非认知能力的重要性及其测量. *中国考试*, (09),22–31.

Agard, C., & von Davier, A. (2018). *The virtual world and reality of testing: Building virtual assessments.* In H. Jiao & R. Lissitz (Eds.), Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective (pp. 1–30). Charlotte, NC: Information Age Publishing.

Autor, D., & Dorn, D. (2009). This job is" getting old": Measuring changes in job opportunities using occupational age structure. *American Economic Review, 99*(2), 45-51.

Baker, C., Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. *In Proceedings of the thirty-third annual conference of the cognitive science society* (pp. 2469–2474).

Banfield, J., & Wilkerson, B. (2014). Increasing student intrinsic motivation and self-efficacy through gamification pedagogy. *Contemporary Issues in Education Research*, *7*(4), 291–298.

Bergner, Y., & von Davier, A. (2018). Process data in NEAP: Past, present, and future. *Journal of Educational and Behavioral Statistics.* doi:10.3102/1076998618784700

Bergner, Y., Shu, Z., & von Davier, A. A. (2014). Visualization and confirmatory clustering of sequence data from a simulation-based assessment task. *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 177–184).

Bezirhan, U., Davier, M. V., & Grabovsky, I. (2021). Modeling item revisit behavior: the hierarchical speed–accuracy–revisits model. *Educational and Psychological Measurement.* doi:10.1177/0013164420950556

Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. Alexandria, VA: ASCD.

Carroll, Kathleen A. & Harris, Carolynn M. (2020). Using a repetitive instructional intervention to improve students' higher-order thinking skills. *College Teaching*, 1–9.

Chen, Y. (2020). A continuous-time dynamic choice measurement model for problem-solving process data. *Psychometrika*, *85*(4), 1052-1075.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333–353.

Dicerbo, K. E. & Kidwai, K. (2013). Detecting player goals from game log files. *Vox Sanguinis*, *11*(3), 350–376.

Diehl, M., Marsiske, M., Horgas, A., Risenberg, A., Saczynski, J., & Willi, S. (2005). The revised observed tasks of daily living: a performance-based assessment of everyday problem solving in older adults. *Journal of Applied Gerontology the Official Journal of the Southern Gerontological Society, 24*(3), 211.

Diserens, D., Schwartz, M. W., Guenin, M., & Taylor, L. A. (1986). Measuring the problem-solving ability of students and residents by microcomputer. *Journal of medical education*, *61*(6), 461-466.

Doerner, D. (1980). On the difficulties people have in dealing with complexity. *Simulation & Gaming*, *11*(1), 87–106.

Fayyad, U., Piatetskyshapiro, G., & Smyth, P. (1996). Knowledge discovery and data mining: Towards a unifying framework. *Shapiro*, 82-88.

Fossey, W. A. (2017). *An evaluation of clustering algorithms for modeling game-based assessment work processes*. Unpublished doctoral dissertation, University of Maryland, College Park. URL https://drum.lib.umd.edu/bitstream/handle/1903/20363/Fossey_umd_0117E_18587.pdf?sequence=1

Fox, J. P. & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research*, *51*(4), 540–553.

Frensch, PA., and Funke, J. (2002). *Thinking and problem solving*. In Psychology, from Encyclopedia of Life Support Systems. (EOLSS), Developed under the Auspices of the UNESCO, edited by N. Cowan. Oxford, UK: Eolss Publishers.

Funke, J. (1983). Einige bemerkungen zu problemen der problemlo seforschung oder: Ist testintelligenz doch ein pra diktor? [Some comments to problems of problem solving research, or: An intelligence test is a predictor, isn't it?]. *Diagnostica*, *29*, 283-302.

Greiff, S., Wüstenberg, S., Holt, D. V., Goldhammer, F., & Funke, J. (2013). Computer-based assessment of complex problem solving: Concept, implementation, and application. *Educational Technology Research and Development*, *61*(3), 407-421.

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education*, *91*, 92–105.

Greiff, S., Wüstenberg, S. & Funke, J. (2012). Dynamic problem solving: A new assessment perspective. *Applied Psychological Measurement*, *36*(3), 189-213.

Han, Z., He, Q., & von Davier, M. (2019). Predictive feature generation and selection using process data from PISA interactive problem-solving items: An application of random forests. *Front Psychology*, *10*, 2461.

Hao, J., Shu, Z., & von Davier, A. (2015). Analyzing process data from game/scenario-based tasks: An edit distance approach. *Journal of Educational Data Mining, 7*(1), 33–50.

Hao, J., Smith, L., Mislevy, R., von Davier, A., & Bauer, M. (2016). *Taming log files from game/simulation-based assessment: Data models and data analysis tools* (Research Report No. RR-16-10). Princeton, NJ: Educational Testing Service.

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with n-grams: Insights from a computer-based large-scale assessment. In A. L. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & S.-M. Chow (Eds.), *Quantitative psychology research. The 79th annual meeting of the psychometric society*, Madison,Wisconsin, 2014 (pp. 750–777).

He, Q., Borgonovi, F., & Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: Using sequence mining to identify behavioral patterns across digital tasks. *Computers & Education, 166.*

Jeon, M., Boeck, P. D., Luo, J., Li, X., & Lu, Z. L. (2021). Modeling within-item dependencies in parallel data on test responses and brain activation. *Psychometrika*, *86*(1).

Jiao, H., & Lissitz, R. (2018). *Technology enhanced innovative assessment development, modeling, and scoring from an interdisciplinary perspective*. Charlotte, NC: Information Age Publishing.

Jiao, H., Liao, D., & Zhan, P. (2019). Utilizing process data for cognitive diagnosis. In M. von Davier & Y.S. Lee (Eds.), *Handbook of Diagnostic Classification Models: Models and Model Extensions, Applications, Software Packages* (pp. 421-436). Cham: Springer International Publishing.

Johnson, R. B. & Christensen, L. (2014). *Educational research: Quantitative, qualitative and mixed methods approaches*, *5th edition*(pp.59-65). Thousand Oaks, CA: SAGE Publications.

Li, J., Zhang, B., Du, H., Zhu, Z., & Li, Y. (2015). Metacognitive planning: Development and validation of an online measure. *Psychological Assessment, 27*(1), 260–271.

Liu, C., & Cheng, Y. (2018). An application of the support vector machine for attribute-by-attribute classification in cognitive diagnosis. *Applied Psychological Measurement, 42*(1), 58–72.

Liu, H., Liu Y., & Li, M. (2018). Analysis of process data of PISA 2012 computer-based problem solving: Application of the modified multilevel mixture IRT model. *Frontiers in Psychology, 9*, 1372.

Man, K., Harring, J. R., Jiao, H., & Zhan, P. (2019). Joint modeling of compensatory multidimensional item responses and response times. *Applied Psychological Measurement, 43*(8).

Man, K., & Harring, J. R. (2020). Assessing pre-knowledge cheating via innovative measures: A multiple-group analysis of jointly modeling item responses, response times, and visual fixation counts. *Educational and Psychological Measurement, 81(*3).

Marshall, J. (1977). Assessment of problem-solving ability. *Medical Education, 11*(5), 329-334.

Mayer, R.E. (1990). *"Problem solving"*, In M. W. Eysenck (Ed.), The Blackwell Dictionary of Cognitive Psychology, Basil Blackwell, Oxford, pp. 284-288.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives, 1*(1), 3–62.

Molenaar, D., Bolsinova, M., & Vermunt, J. (2018). A semi-parametric within-subject mixture approach to the analyses of responses and response times. *British Journal of Mathematical and Statistical Psychology, 71*(2), 205-228.

Molenaar, D., Oberski, D., Vermunt, J., & De Boeck, P. (2016). Hidden markov item response theory models for responses and response times. *Multivariate behavioral research, 51*(5), 606-626.

NCES. (2014). *NAEP TEL Wells sample item*. National Center for Education Statistics. Retrieved February 24, 2019, from http://nces.ed.gov/nationsreportcard/tel/wells_item.aspx

Novak, J. D. (1961). An approach to the interpretation and measurement of problem solving ability. *Science Education, 45*(2), 122-131.

OECD (2003). *The PISA 2003 assessment framework: Mathematics, reading, science and problem solving knowledge and skills.* Paris: OECD Publishing.

OECD (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Paris: OECD Publishing.

OECD (2014). *PISA 2012 results: Creative problem solving: Students' skills in tackling real-life problems (Volume V)*. Paris: OECD Publishing.

OECD (2016). *PISA 2015 assessment and analytical framework: Science, reading, mathematic and financial literacy.* Paris: PISA, OECD Publishing.

OECD (2019). *PISA 2021 creative thinking framework: Third Draft[R]*. Paris: OECD Publishing."

Omodei, M. M., & Wearing, A. J. (1995). The fire chief microworld generating program: an illustration of computer-simulated microworlds as an experimental paradigm for studying complex decision-making behavior. *Behavior Research Methods Instruments & Computers, 27*(3), 303-316.

Qiao, X., & Jiao, H. (2018). Data mining techniques in analyzing process data: A didactic. *Frontiers in Psychology, 9*, 2231.

Shute, V., & Moore, G. (2018). Consistency and validity in game-based stealth assessment. In H. Jiao & R. Lissitz (Eds.), *Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective* (pp. 31–51). Charlotte, NC: Information Age Publishing.

Shute, V. J., & Rahimi, S. (2020). Stealth assessment of creativity in a physics video game. *Computers in Human Behavior, 116*, 1–13.

Shute, V., Ke, F., & Wang, L. (2017). Assessment and adaptation in games. In P. Wouters & H. van Oostendorp (Eds.), *Instructional techniques to facilitate learning and motivation of serious games* (pp. 59–78). New York, NY: Springer.

Shu, Z., Bergner, Y., Zhu, M., Hao, J., von Davier, A. (2017). An item response theory analysis of problem-solving processes in scenario-based tasks. *Psychological Test and Assessment Modeling, 59*(1), 109–131.

Soller, A., & Stevens, R. (2007). Applications of stochastic analyses for collaborative learning and cognitive assessment. In G.R. Hancock & K.M. Samuelsen (Eds.) *Advances in Latent Variable Mixture Models*, pp. 217–253. Information Age Publishing.

Stanek, S. & Sabat, A. (2019). The use of IT tools in the assessment and development of leadership abilities. *Problemy Zarządzania - Management Issues, 5*(85), 89-110.

Ulitzsch, E., He, Q., Ulitzsch, V., Molter, H., & Pohl, S. (2021). Combining clickstream analyses and graph-modeled data clustering for identifying common response processes. *Psychometrika, 86*(1).

Unal, E., & Cakir, H. (2021). The effect of technology-supported collaborative problem solving method on students' achievement and engagement. *Education and Information Technologies, 26*, 4127-4150.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*(2), 181-204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*(3), 287-308.

Wang, S. Y., Zhang, S. S., Douglas, J., & Culpepper, S. (2018). Using response times to assess learning progress: A joint model. for responses and response times. *Measurement: Interdisciplinary Research and Perspectives, 16*(1), 45- 58.

Weir, K. (2018). *Designing smarter tech tools: New technology in educational gaming, health-care communication, robotics and more is benefiting from psychologists' input.* URL https://www.apa.org/monitor/2018/11/cover-tech-tools.aspx

Zhan, P., & He, K. (2021). A longitudinal diagnostic model with hierarchical learning trajectories. *Educational Measurement: Issues and Practice.* https://doi.org/10.1111/emip.12422

Zhan, P., & Qiao, X. (2020, July 13). A diagnostic classification analysis of problem-solving competence using process data. https://doi.org/10.31234/osf.io/wtyae

Zhan, P., Jiao, H., & Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology, 71*(2), 262–286.

Zhao, W., Shute, V., & Wang, L. (2015). Stealth assessment of problem-solving skills from gameplay. *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*, (15212), 1–11.

Zoanetti, N. (2010). Interactive computer-based assessment tasks: How problem-solving process data can inform instruction. *Australasian Journal of Educational Technology, 26*, 585-606.

The measurement of problem-solving competence using process data

LIU Yaohui[1], XU Huiying[1], CHEN Qipeng[1], ZHAN Peida[1, 2]

([1]*Department of Psychology, College of Teacher Education, Zhejiang Normal University, Jinhua, 321004, China)*

([2]*Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, Zhejiang Normal University, Jinhua, 321004, China)*

Abstract: Problem-solving competence is an individual's capacity to engage in cognitive processing to understand and resolve problem situations where a method of solution is not immediately obvious. The measurement of problem-solving competence requires the use of relatively more complex and real problem situations to induce the presentation of problem-solving behaviors. This brings challenges to both the measurement methods of problem-solving competence and the corresponding data analysis methods. Using virtual assessments to capture the process data in problem-solving and mining the potential information contained therein is a new trend in measuring problem-solving competence in psychometrics. To begin with, we reviewed the development of the measurement methods of problem-solving competence: from paper-and-pencil tests to virtual assessments. In addition, we summarized two types of process data analysis methods: data mining and statistical modeling. Finally, we look forward to possible future research directions from five perspectives: the influence of non-cognitive factors on problem-solving competence, the use of multimodal data to measure problem-solving competence, the measurement of the development of problem-solving competence, the measurement of other higher-order thinking competencies, and the definition of concept and structure of problem-solving competence.

Key words: problem-solving competence, processing data, virtual assessment, computer-based assessment, higher-order thinking competence

附录：

Problem Solving Ability

first [1] answer

1. It is known that general ability is important to success in any course. If the students in the experimental class had higher general ability than those in the control class, we migh expect that they would receive higher final grades in physics than the control students.
2. Instruction in reading has been found to improve the reading comprehension of students in some cases. We should expect that instruction in reading the physics textbook would result in higher final grades for the students in the experimental class.

* * * * * * * * * *

above answer [1|1] plus a second

1

1. Mr. K. studied the American Council on Education (ACE) exam scores for the students in the experimental and control groups. He found that the average scores were about equal. This was important for him to know in order that he could proceed to make comparisons of the experimental and control groups.
2. Mr. K. obtained the high school percentile ranks (HPR) for each of his students. He found that the students of the control and the experimental groups had about the same high school ranks, on the average. This suggests that no differences are to be found between the groups when the final grades are compared.

2

1. Many textbooks contain more information than is important. By giving instruction to the experimental group as to what material should be read most thoroughly, we could expect that Mr. K.'s experimental group would get higher final grades.
2. Many students taking science courses are poor readers. The experimental group should have a definite advantage over the control group, if they are taught how to read the textbook.

* * * * * * * * * *

above answer [1|1|1] plus last choice

1–1

1. Mr. K. found that the students in the control group received about as good a grade on the final exam as did students in the experimental group. One should conclude that instruction in how to read a textbook does not improve a student's ability to do well in that course.
2. Mr. K. compared grades in the course for students in the experimental and control groups. Since the students did about equally well on the American Council on Education Exam when they were compared, we should not expect to find a difference in their course grades.

1–2

1. When Mr. K. made a statistical analysis of the differences between the grades received by the control group and the experimental group, he could not find any statistically significant differences. He should have expected this result, since the students in the two groups had about equal high school ranks.
2. The fact that Mr. K. could find no statistical difference between the control and the experimental group's grades illustrates the weakness of statistics. Perhaps he would have done better to ask his colleagues to study the two sets of grades and decide whether or not they appeared to be different, on the average.

2–1

1. Most science textbooks contain many scientific terms. Help in understanding these terms should have resulted in higher grades for the experimental group.
2. Physics textbooks often have many graphs and charts. If the instructor helped to interpret these, there is a good chance that the experimental group would get better grades than the control group.

2-2

1. Ability to concentrate on the material being read has been shown to result in higher reading comprehension. Students should be able to concentrate on their reading if they are given instruction in textbook reading, and consequently they should get higher grades.
2. There is some evidence that fast readers are also better readers. If the instructor points out how to best read a chapter, the students can read it faster and therefore better. This would be a good reason why students with instruction in textbook reading might get better grades than students without such instruction.

图 A1. 问题解决能力测试例题(Novak, 1961).



图 A2. 植物大战僵尸游戏截屏(Zhao et al., 2015).

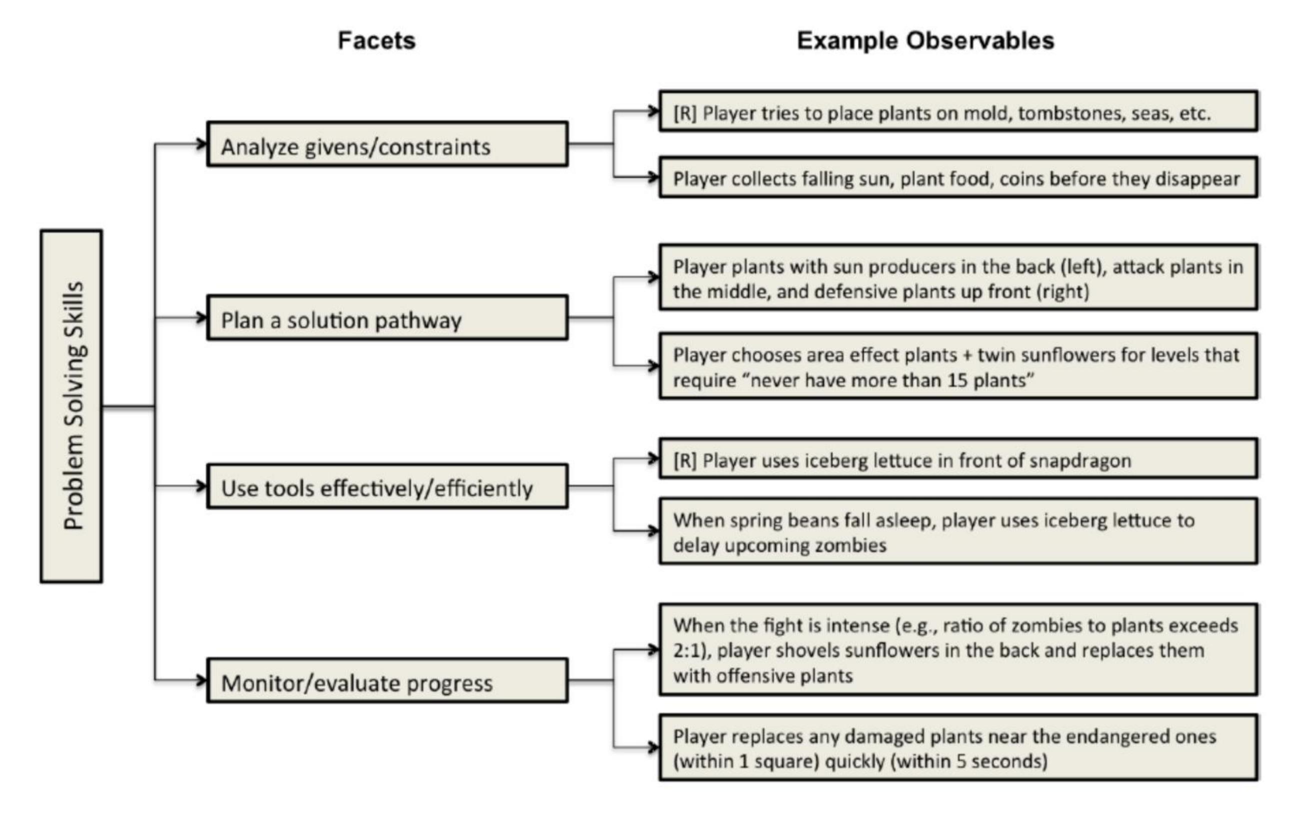
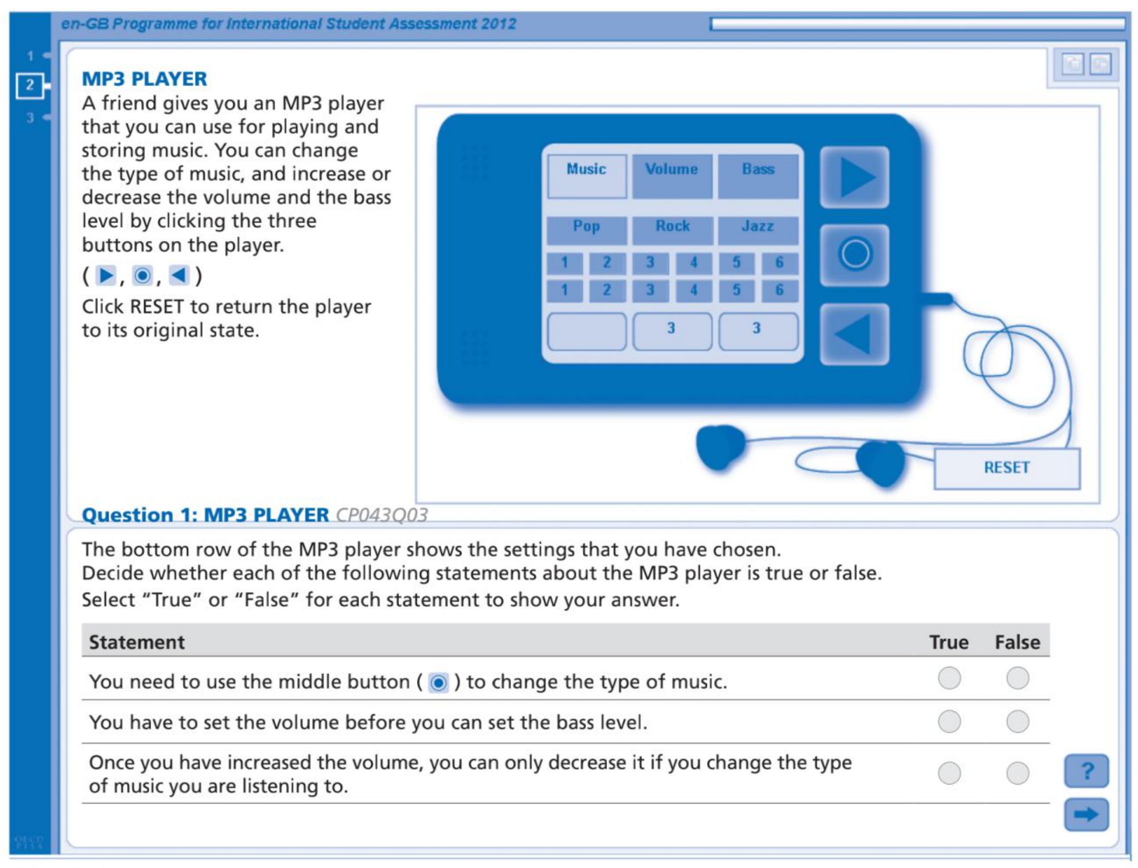


图 A3. 能力模型和一些行为指标之间的联系(Zhao et al., 2015).



QUESTION 2

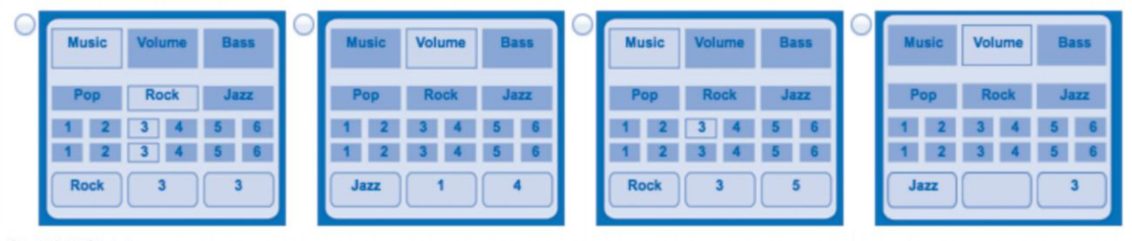
Set the MP3 player to Rock, Volume 4, Bass 2.

Do this using as few clicks as possible. There is no RESET button.

QUESTION 3

Shown below are four pictures of the MP3 player's screen. Three of the screens cannot happen if the MP3 player is working properly. The remaining screen shows the MP3 player when it is working properly.

Which screen shows the MP3 player working properly?



QUESTION 4

Describe in the box below how you could change the way the MP3 player works so that there is no need to have the bottom button (◀).

You must still be able to change the type of music, and increase or decrease the volume and the bass level.



图 A4. PISA 2012 问题解决测试例题.